Università degli Studi del Molise
Dipartimento di Economia, Gestione, Società e Istituzioni

ECONOMICS & STATISTICS DISCUSSION PAPER
No. 073/14

Identifying I(0) Series in Macro-panels: Are Sequential Panel Selection Methods Useful?

Mauro Costantini

Claudio Lupi

Identifying I(0) Series in Macro-panels: Are Sequential Panel Selection Methods Useful?*

Mauro Costantini† and Claudio Lupi⋆
† *Department of Economics and Finance – Brunel University. Kingston Lane, Uxbridge, Middlesex UB8 3PH (UK). E-mail: Mauro.Costantini@brunel.ac.uk*
⋆ *Department of Economics – University of Molise. Via F. De Sanctis, I-86100 Campobasso (Italy). E-mail: lupi@unimol.it*

Abstract
Sequential panel selection methods (SPSMs) are based on the repeated application of panel unit root tests and are increasingly used to identify $I(0)$ time series in macro-panels. We check the reliability of SPSMs by using Monte Carlo simulations based on generating the individual test statistics and the p values to be combined into panel unit root tests, both under the unit root null and under selected local alternatives. The analysis is carried out considering both independent and dependent test statistics. We show that SPSMs do not possess better classification performances than conventional univariate tests.

1. Introduction

Panel unit root (UR) tests are powerful tools to check the null hypothesis that all the time series in a panel are $I(1)$. However, they are unsuitable to classify individual time series into nonstationary and stationary ones. In order to overcome this shortcoming, Chortareas & Kapetanios (2009) proposed using a sequential panel selection method (SPSM) based on the application of a panel UR test recursively, at each iteration eliminating from the panel the single series for which individual evidence in favour of the stationary hypothesis is strongest. This procedure was initially conceived using individual Dickey-Fuller (DF) tests jointly with the panel unit root test developed by Im *et al.* (2003). However, different flavours of the SPSM can be obtained using different tests.

Although the procedure has received increasing interest,[1] not much is known about its real merits as a classification tool, the only evaluation being limited to the experiments carried out in Chortareas & Kapetanios (2009).

In this paper we investigate the performance of the SPSM as a classification device. We complement and extend Chortareas & Kapetanios's (2009) analysis along five directions:

(i) we study the performance of the procedure using four different panel UR tests. Of course, the classification performance of each SPSM may depend on the characteristics of the selected UR test. In this paper we consider the tests developed in Im *et al.* (2003), Choi (2001), Demetrescu *et al.* (2006), and Hanck (2013). We label the corresponding SPSMs as I-SPSM, C-SPSM, D-SPSM, and H-SPSM, respectively. For reasons explained below, all the panel UR tests used in this paper are built upon univariate Dickey-Fuller (DF) tests. The classification performance of the SPSMs is also compared to the results

* Preliminary versions of this paper were presented at the Economics Seminars at the Department of Economics and Finance, Brunel University and at the 7th International Conference on Computational and Financial Econometrics (CFE 2013). We are grateful to all participants for their comments and suggestions. We owe special thanks to Gaia Becheri and Xuguang Sheng for discussion.
JEL classification numbers: C12, C15, C23.
Key words: Unit root, Panel data, Simulation, ROC curve, Discrete classifier.

[1] At the time of writing (January 2014), Chortareas & Kapetanios's (2009) paper has received more than 60 citations.

obtained by applying individual DF tests as well as Hommel's (1988) multiple testing procedure;

(ii) we use local-to-unit root alternatives (Phillips, 1987), rather than fixed alternatives as done in Chortareas & Kapetanios (2009). This choice allows us to treat the test statistics and the p values under the null and under the alternative in a consistent way;

(iii) we analyze the effect of cross-section dependence on the classification ability of the procedure. This aspect is of crucial practical importance, given that typical macro-panels can be expected to be cross-sectionally correlated (see, e.g., O'Connell, 1998);

(iv) we focus on the classification performance of the procedure in a way which is not influenced by the finite-sample performance of the underlying individual DF tests. In practice, our simulation results can be read as quality upper bounds of the results that can be attained by using SPSMs in practical circumstances;

(v) simulation results are illustrated using Receiver Operating Characteristic (ROC) graphs, consistently with the literature on discrete classifiers (see, e.g., Fawcett, 2006).

The analysis is carried out by using extensive Monte Carlo experiments. Since we want to isolate the properties of the sequential procedure from those of the underlying individual and panel tests, rather than simulating the individual time series, we simulate the asymptotic (with $T \to \infty$) DF individual t statistics and p values under the UR null and under selected local-to-unit root alternatives. These simulated values constitute the input for the different classification procedures. In so doing, the classification results depend only on the procedure, not on the specification and the finite-sample properties of the underlying DF tests. The asymptotic power of the individual DF tests can be varied by varying the distance between the null and the (local) alternative hypothesis.

The remainder of this paper is organized as follows: Section 2 describes the panel tests considered in the paper and draws some implications for their use jointly with the SPSM; Section 3 is devoted to the description of the Monte Carlo experiments; Section 4 discusses the simulation results, and Section 5 concludes.

2. Panel unit root tests

In our opinion there are some classes of panel UR tests that are better suited than others to be applied jointly with the SPSM. In particular, it is necessary to consider panel UR tests whose null hypothesis is that all the series are $I(1)$, whereas the alternative hypothesis must be that at least one of the series is $I(0)$. Furthermore, the selected panel test should preferably be built by aggregating the results of individual time series tests, so that the resulting panel test is consistent with the selection criterion used to eliminate from the panel one series at each iteration. Finally, given that the selected panel test is going to be applied over a decreasing number N of series, panel tests that do not rely on N-asymptotics should be preferred. All these requirements drastically reduce the number of suitable panel UR tests and we argue that, beside Im *et al.*'s (2003) test, p value combination tests (see, e.g., Choi, 2001; Demetrescu *et al.*, 2006) and Hanck's (2013) intersection test are natural candidates to be used jointly with the SPSM. Although in this paper we adhere to the panel UR tests based on individual DF tests, as far as the p value combination and intersection tests are considered, different individual unit root tests could be utilized as well (see, e.g., Costantini & Lupi, 2013). Finally, since Hanck's (2013) approach is based on Simes's (1986) test, and given that Hommel's (1988) procedure is equivalent to applying Simes's (1986) test to each intersection hypothesis of a closed testing procedure, the classification performance of the multiple testing approach proposed by Hommel (1988) is also considered in our Monte Carlo analysis.

The mean-t test

Im *et al.* (2003, p. 60) showed that for fixed N and T, under the joint UR null the distribution of the average of N independent individual DF t statistics

$$\bar{t}_{N,T} := \frac{1}{N} \sum_{i=1}^{N} t_{i,T} \,. \tag{1}$$

is non-standard but does not depend on nuisance parameters. The test built on eq. (1) rejects the null for large (in absolute value) negative values of the $\bar{t}_{N,T}$ statistic. The asymptotic (in T) distribution of the test statistic shifts to the right as N increases, so that the critical values become “less negative” for large values of N. This fact, together with the observation that under the alternative hypothesis the test statistic diverges, makes it clear that the power of the test increases with N and with the number of $I(0)$ series in the panel. However, the mean-t test may be heavily biased in the presence of cross-dependent series (see, e.g., O'Connell, 1998; Maddala & Wu, 1999).

The p value combination test

Choi (2001) proved that under mild regularity conditions, if $N < \infty$ time series in a panel are cross-sectionally independent, then under the joint UR null

$$Z_c := \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \Phi^{-1}(p_i) \xrightarrow{w} \mathrm{N}(0,1) \tag{2}$$

as $T \to \infty$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function (CDF) and p_i $(i = 1, \ldots, N)$ are the individual p values from N ordinary DF tests carried out on each series in the panel. The test is one-sided and the null is rejected for $Z_c < \Phi^{-1}(\alpha)$ for a given significance level α.

Choi's (2001) test has high power: in the presence of, say, $N = 10$ individual p values all around 0.3, no UR test would individually reject the null, but the panel test would reject the global null hypothesis that all the series are $I(1)$ at the 5% significance level. In general, eq. (2) implies that the panel test will reject in the presence of N equal p values smaller than $\Phi\left(\Phi^{-1}(\alpha)/\sqrt{N}\right)$, which (for the usual significance levels α) is increasing in N. From eq. (2) it is also clear that for the panel test to reject it is sufficient that the average probit $\bar{t} := N^{-1} \sum_{i=1}^{N} \Phi^{-1}(p_i)$ be smaller than $N^{-\frac{1}{2}}\Phi^{-1}(\alpha)$, which is decreasing in absolute value with N. Furthermore, the density of the p values under the alternative is markedly skewed to the right, with the skewness increasing with the distance of the alternative from the null. This makes rejection more likely the larger is the number of hypotheses under the alternative and the larger is their distance from the null. This in turn means that for a given set of alternatives the power of the panel UR test increases with the fraction of hypotheses under the alternative and their distance from the null.

Choi's (2001) test can be extended to cover the case of cross-dependent time series, as suggested in Demetrescu *et al.* (2006).

The intersection test

The panel UR intersection test proposed in Hanck (2013) is based on the testing procedure developed in Simes (1986) and tests the global hypothesis that all the series have a unit root, while controlling for the probability of falsely rejecting at least one individual true null hypothesis at the chosen significance level α. By controlling the family-wise error rate, the intersection test tends to be more conservative than other tests. Indeed, Hanck (2013) showed that the intersection test has lower power than those proposed in Maddala & Wu (1999) and Demetrescu *et al.* (2006) in many circumstances.

Contrary to what happens with the previously discussed tests, the number of series under the alternative hypothesis has no direct influence on power. In fact, it is sufficient a single small p value for the intersection test to reject the null, whereas the presence of large p values does not influence the result (there is no averaging).

Being based on Simes's (1986) procedure, the intersection test can be used in the presence of positively dependent tests (Sarkar & Chang, 1997).

Implications for SPSMs

Following Chortareas & Kapetanios's (2009) argument, the case for SPSMs is stronger the larger is the power gain of the panel UR tests over the individual time series ones. Therefore, as far as the I-SPSM, the C-SPSM, and the D-SPSM are concerned, a few implications can be advanced:

(i) The case for the I-SPSM, the C-SPSM, and the D-SPSM is stronger when N is fairly large and most of the series are $I(0)$.

(ii) The performance of the I-SPSM, of the C-SPSM, and of the D-SPSM should deteriorate as the sequence of panel UR tests proceeds, since both N and the proportion of series under the alternative decrease at each step.

Implication (i) was already suggested by Chortareas & Kapetanios (2009) with reference to the I-SPSM. However, the fact that panel UR tests are more powerful when the fraction of stationary series is large may be somewhat disturbing for macroeconomists who are often more interested in the opposite situation where most of the series in a panel are $I(1)$ and only a few are $I(0)$.

Panel test power is not the only element that influences the ability of SPSMs to correctly classify time series. In fact, in order for SPSMs to correctly classify a $I(0)$ time series the panel test should reject the null and the minimum individual p value, $p_{\min}$, should correspond to a truly $I(0)$ series. While this happens with probability 1 asymptotically when fixed alternatives are considered, this is not guaranteed either in finite samples or asymptotically with local alternatives. Indeed, the probability that a p value as small or smaller than $p_{\min}$ corresponds to a test under the alternative is

$$\begin{aligned} \Pr(H_1 | p \leq p_{\min}) &= \frac{\Pr(p \leq p_{\min} | H_1) \Pr(H_1)}{\Pr(p \leq p_{\min} | H_1) \Pr(H_1) + \Pr(p \leq p_{\min} | H_0) \Pr(H_0)} \\ &= \frac{F_1(p_{\min}) \Pr(H_1)}{F_1(p_{\min}) \Pr(H_1) + p_{\min}(1 - \Pr(H_1))} \end{aligned} \tag{3}$$

where $F_1(\cdot)$ is the CDF of the p values of the individual UR tests under the alternative which is determined by the test used and by the distance between the null and the local alternative hypothesis. From eq. (3) it can be verified that

$$\frac{\partial \Pr(H_1 | p \leq p_{\min})}{\partial \Pr(H_1)} = \frac{p_{\min} F_1(p_{\min})}{D^2} > 0 \tag{4}$$

and

$$\frac{\partial \Pr(H_1 | p \leq p_{\min})}{\partial F_1(p_{\min})} = \frac{p_{\min} \Pr(H_1) \left[1 - \Pr(H_1)\right]}{D^2} > 0 \tag{5}$$

with D^2 the squared denominator of eq. (3). In other words, denoting by N_1 the number of tests (or series) under the alternative, the probability that a p value as small or smaller than $p_{\min}$ corresponds to a hypothesis under H_1 increases with $\Pr(H_1) = N_1/N$ as well as with the distance between the null and the local alternative.

Following a similar line of reasoning for $\Pr(H_0|p \leq p_{\min})$ makes it possible to write the odds

$$\Omega_{10} := \frac{\Pr(H_1|p \leq p_{\min})}{\Pr(H_0|p \leq p_{\min})} = \frac{\Pr(p \leq p_{\min}|H_1)\Pr(H_1)}{\Pr(p \leq p_{\min}|H_0)\Pr(H_0)} = \frac{F_1(p_{\min})}{p_{\min}} \times \frac{N_1/N}{1 - N_1/N} \tag{6}$$

from which it is clear that $\partial\Omega_{10}/\partial(N_1/N) > 0$. Of course, for a given ratio N_1/N,

$$\Omega_{10} \propto \frac{F_1(p_{\min})}{p_{\min}}. \tag{7}$$

Since the skewness of the density of the p values under the alternative increases with the distance between the null and the alternative, the odds (6) also increase with the distance between the null and alternative. These results have important implications as far as the classification procedure is concerned. Indeed, eq. (6) implies a deterioration of the procedure's performance as the iteration process proceeds that goes beyond the effect related to the power of the panel UR test. Furthermore, for a given value of the ratio N_1/N, mean-t and p value combination tests will tend to reject the global null more easily for increasing values of N. However, if the alternative is sufficiently close to the null, the odds ratio (7) indicates that there is a non trivial probability that a truly null individual hypothesis is rejected in the iterative procedure, and this phenomenon will tend to increase with N, given that the panel test will be able to reject the global null in the presence of higher individual p values (smaller, in absolute value, test statistics). On the other hand, when the null and the alternative hypotheses are well separated, this effect reduces and it is more likely that the individual hypothesis rejected in the iterative procedure is indeed correctly rejected. However, as far as the null and the alternative hypotheses are well separated, the power of conventional univariate UR tests is satisfactory, rendering the application of SPSMS potentially superfluous.

3. Monte Carlo design

Six different classification procedures are compared, namely the I-SPSM, the C-SPSM, the D-SPSM, the H-SPSM, the univariate DF, and Hommel's (1988) method.

The null hypothesis is that all the N series in the panel are generated by

$$y_t = \varrho y_{t-1} + \epsilon_t \tag{8}$$

with $\varrho = 1$ whereas under the alternative a fraction N_1/N of series in the panel are generated by eq. (8) with

$$\varrho = \exp\left(-\frac{\gamma}{T}\right) \approx 1 - \frac{\gamma}{T} \tag{9}$$

with $\gamma > 0$. In our simulations we considered four distinct local-to-unit root alternatives ($\gamma \in \{1, 5, 10, 20\}$).

All experiments have been simulated over 5,000 replications taking the nominal significance level $\alpha = 0.05$, the number of time series $N \in \{10, 20, 40, 80\}$, and the fraction of stationary alternatives $N_1/N \in \{0.20, 0.50, 0.80\}$. Finally, the correlation among individual test statistics has been taken as $\rho \in \{0, 0.4, 0.8\}$. In order to maintain the paper within a reasonable length, we considered only individual DF tests with drift and no trend. All simulations were carried out using R 3.0.1 (R Development Core Team, 2013).

Simulation of the individual t statistics and p values

The simulation algorithm can be divided into 5 steps:

1. define the correlation matrix among the test statistics as

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho & \rho & \rho & \dots & 1 \end{pmatrix}$$

   with $\rho \in \{0, 0.2, 0.4, 0.8\}$;

2. draw N (with $N \in \{10, 20, 40, 80\}$) values $\boldsymbol{n}$ from the N-variate normal $\mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$;

3. draw N possibly correlated uniform variables $\boldsymbol{u}$ as $\boldsymbol{u} = \Phi(\boldsymbol{n})$ and partition the result as $\boldsymbol{u} = (\boldsymbol{u}_1', \boldsymbol{u}_0')'$, where $\boldsymbol{u}_1$ and $\boldsymbol{u}_0$ have N_1 and N_0 elements, respectively;

4. draw N_1 test statistics under the alternative hypothesis as $s_1 = F_1^{-1}(\boldsymbol{u}_1)$ and N_0 test statistics under the null as $s_0 = F_0^{-1}(\boldsymbol{u}_0)$, where $F_1(\cdot)$ and $F_0(\cdot)$ are the CDFS of the test statistics under the alternative and under the null hypothesis, respectively;

5. generate the p values under the alternative as $F_0(s_1)$ and set the p values under the null simply equal to $\boldsymbol{u}_0$.

Step 4 requires inverting the CDF of the test statistics under the null and under the alternative. Therefore, we simulated 200,000 values of the test statistics under the null and under the alternative and used the empirical CDF's $\hat{F}_0(\cdot)$ and $\hat{F}_1(\cdot)$ in step 4. More specifically, under the null the t statistics are asymptotically distributed according to the DF distribution, so that their simulation rises no special difficulty (see, e.g., Hatanaka, 1996, Chapter 7). Furthermore, Phillips (1987) showed that the asymptotic distribution of the DF t statistics under the local-to-unit root alternative can be written in terms of functionals of the Ornstein-Uhlenbeck process, which we simulated following Chan (1988), method II. Finally, the empirical CDF $\hat{F}_0(\cdot)$ is again used to derive the p values.

In order to apply Im *et al.*'s (2003) test recursively, the critical values of the panel UR test for all the possible values of $N \in \{1, \dots, 80\}$ are needed. In order to derive the critical values for the panel test, we simulated $\bar{t}_{N,\infty}$ under the null over 50,000 replications with $N \in \{1, 2, \dots, 10, 15, 20, \dots, 100, 120, 140, \dots, 200\}$ and computed the 100α-th percentile of each simulated distribution, $cv_{N,\alpha}$. Finally, we estimated a response surface of the form

$$cv_{N,\alpha} = \beta_0 + \beta_1 \ln(N) + \beta_2 \ln(N)^2 + \beta_3 \ln(N)^3 + \xi_{N,\alpha}\,. \tag{10}$$

The approximation we got from the response surface for $\alpha = 0.05$ is excellent, with $R^2 = 0.9998$. The 5% critical values were finally computed as

$$\begin{aligned} \widehat{cv}_{N,0.05} &= \hat{\beta}_0 + \hat{\beta}_1 \ln(N) + \hat{\beta}_2 \ln(N)^2 + \hat{\beta}_3 \ln(N)^3 \\ &\approx -2.851 + 0.597 \ln(N) - 0.108 \ln(N)^2 + 0.007 \ln(N)^3 \end{aligned} \tag{11}$$

for any $N \in \{1, \dots, 200\}$. As a by-product of our analysis, eq. (11) generalizes and extends Table 2 in Im *et al.* (2003).

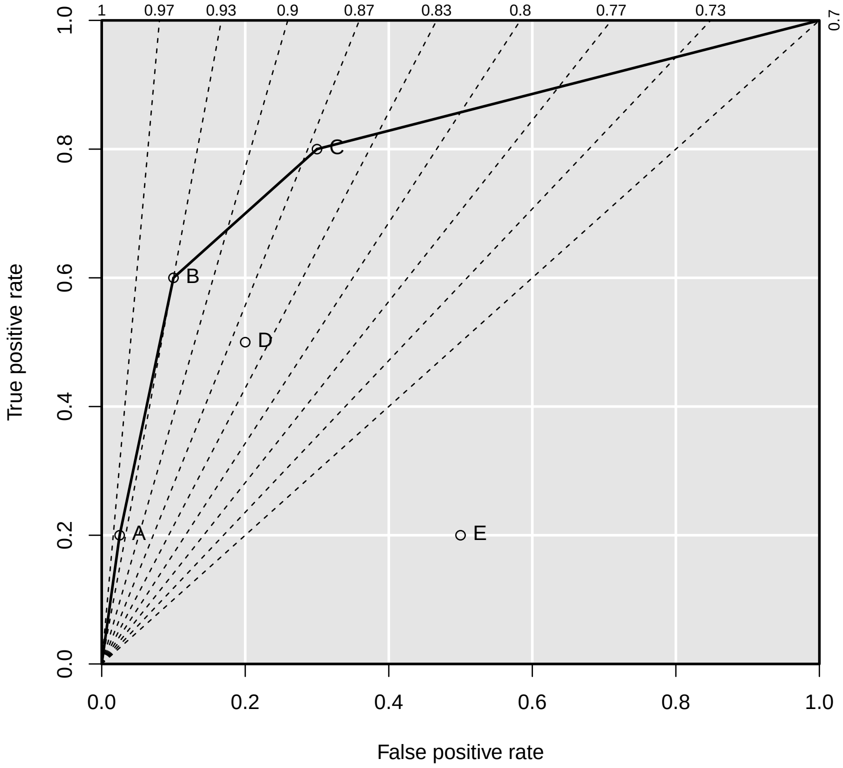

Figure 1. Five discrete classifiers plotted on the ROC space. The dashed lines are precision isometrics (for $N_1/N = 0.7$). The lowest precision isometric coincides with the random guessing line. The broken solid line is the ROC convex hull.

4. Monte Carlo results

The procedures considered in the Monte Carlo simulations can be thought of as different discrete classifiers, and we analyze the simulation results accordingly. In particular, in order to compare the ability of the different procedures to correctly classify the time series into $I(0)$ series (the positives) and $I(1)$ ones (the negatives), we represent the outcome of the Monte Carlo simulations on the Receiver Operating Characteristics (ROC) space (see, e.g., Fawcett, 2006). More precisely, the result of each procedure in each experiment is represented as a point on the ROC space.

A ROC graph is built by plotting the true positive rate (tpr) of a classifier against its false positive rate (fpr), with the tpr and the fpr being defined as

$$\begin{aligned} tpr &:= \frac{\text{number of correctly classified positives}}{\text{total number of true positives}} \\ fpr &:= \frac{\text{number negatives incorrectly classified as positives}}{\text{total number of true negatives}}. \end{aligned}$$

Figure 1 presents an explicating example, where five discrete classifiers are plotted on the ROC space. The diagonal line $tpr = fpr$ represents the situation of randomly guessing the classification. Points below this line (point E, in Figure 1) identify classifiers that perform worse than random guessing. The perfect classifier is represented by the point $(0, 1)$. Precision isometrics (the dashed lines in Figure 1) help identifying those classifiers that have the same performance in terms of precision (Flach, 2003), where precision is defined

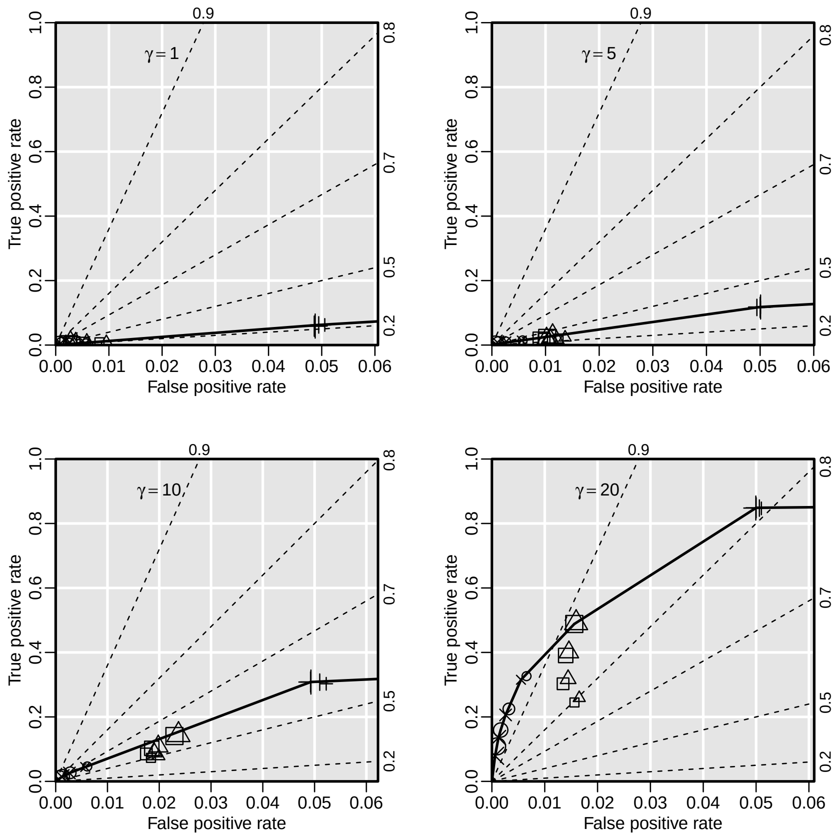

Figure 2. Independent test statistics. 5,000 replications with the fraction of stationary series $N_1/N = 0.2$. The local-to-unit root alternative is $\varrho = \exp(-\gamma/T)$ with $\gamma \in \{1, 5, 10, 20\}$. Larger symbols correspond to larger panels with $N \in \{10, 20, 40, 80\}$. The dashed lines are precision isometrics. The lowest precision isometric coincides with the random guessing line. The broken solid line is the ROC convex hull. $\square$ = C-SPSM; $\triangle$ = I-SPSM; $\bigcirc$ = H-SPSM; $+$ = DF; $\times$ = Hommel.

as

$$prec := \frac{\text{number of correctly classified positives}}{\text{total number of classified positives}} .$$

The level of precision relative to each precision isometric is reported on the border of the graph. The broken line is the ROC convex hull (Provost & Fawcett, 2001), the most "up-left" broken line joining some classifiers on the ROC space. Classifiers lying between the random guessing line and the convex hull (e.g., point D in Figure 1), despite doing better than random guessing, are sub-optimal.

We contend that analyzing the simulation results using the ROC framework is more appropriate than just reporting the estimated probability of correctly classifying a series as stationary. In fact, by using ROC graphs we can jointly examine not only the true positive rate, but also the false positive rate and the precision of each classifier. Furthermore, the graphical representation of the results eases comparisons and makes evident details that could go unnoticed in large tables.

Independent test statistics

To save space, in this Section we report only the simulation results relative to $N_1/N \in \{0.2, 0.8\}$: the results with $N_1/N = 0.5$ are qualitatively similar and are available in the supplementary material reported in the appendix.

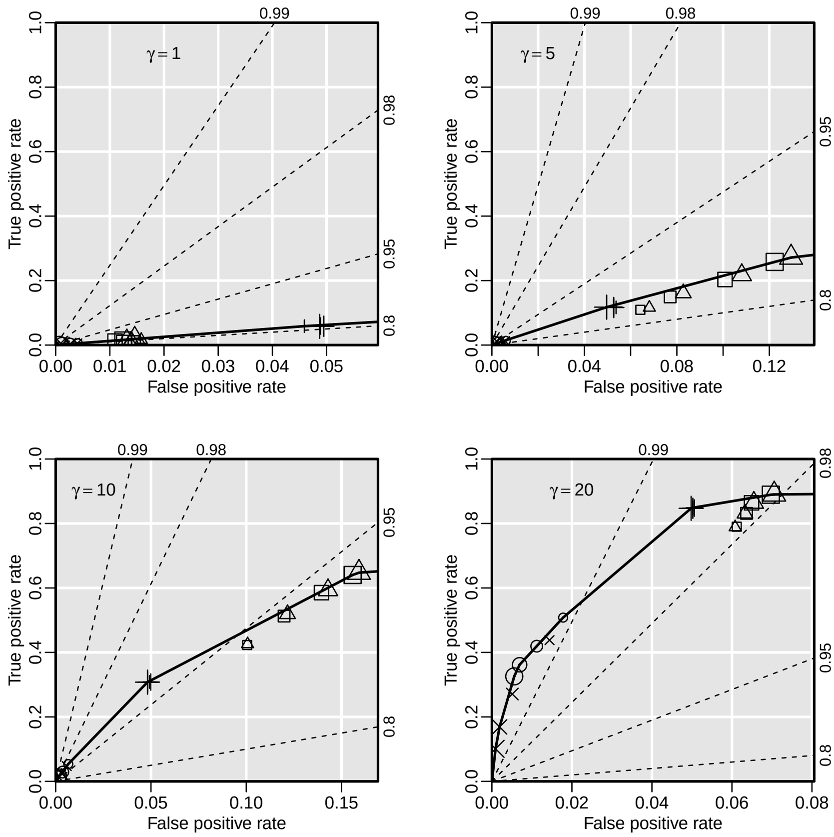

Figure 3. Independent test statistics. 5,000 replications with the fraction of stationary series $N_1/N = 0.8$. The local-to-unit root alternative is $\varrho = \exp(-\gamma/T)$ with $\gamma \in \{1, 5, 10, 20\}$. Larger symbols correspond to larger panels with $N \in \{10, 20, 40, 80\}$. The dashed lines are precision isometrics. The lowest precision isometric coincides with the random guessing line. The broken solid line is the ROC convex hull. $\square$ = C-SPSM; $\triangle$ = I-SPSM; $\bigcirc$ = H-SPSM; $+$ = DF; $\times$ = Hommel.

The I-SPSM and C-SPSM procedures have very similar performances. When the fraction of stationary time series is small (see Figure 2), either procedure does not provide advantages over the standard time series approach, as predicted. Quite on the contrary, irrespective of the "distance" of the local alternative from the UR null hypothesis, these two SPSM-based procedures tend to be too conservative and to be sub-optimal for small N. When $\gamma = 20$, the predicted effect of panel dimension (N) on the classification performance is evident. The H-SPSM and Hommel's method are very conservative and of scarce practical use under these conditions. In a classification exercise, control of the family-wise error rate implied by the two procedures is probably excessive, giving rise to an overall weak classification criterion. Although expected, these results are nevertheless quite disappointing given that it is easy to imagine situations in which the main interest lies in distinguishing a few $I(0)$ time series in a panel made of mostly nonstationary macroeconomic time series. In such a setting the SPSM will not be of great help, whereas the conventional approach related to individual UR testing ensures better results.

When most of the series in the panel are $I(0)$ (see Figure 3) the results are more articulated. When the alternative hypothesis is very close to the null (as is the case when $\gamma = 1$), then no method is significantly better than random guessing (in fact, the same is true also when the fraction of tests under the alternative is small). When the null and the alternative hypotheses are more separated (e.g., when $\gamma = 5$ or $\gamma = 10$), both the true positive rate and the false positive rate of the I-SPSM and of the C-SPSM increase with N. In particular, the false positive rate exceeds 0.15 for $\gamma = 10$ and $N > 40$. In this respect, our results square nicely with those reported in Chortareas & Kapetanios (2009, Table 2). Since Chortareas & Kapetanios (2009) used a fixed alternative where on average $\varrho = 0.85$, the most appropriate comparison is between their results relative to the I-SPSM with $N_1/N = 0.80$ ($I(1)\% = 0.20$, in their notation) and $T = 50$, with our results relative to the same procedure with $N_1/N = 0.80$ and $\gamma = 10$ (the exactly corresponding value of γ would be 7.5). The results are indeed quite similar. However, given that Chortareas & Kapetanios (2009) focused on fixed alternatives, in their simulations the performance of the I-SPSM improves substantially for larger values of T. On the contrary, we show that, when appropriate local alternatives are considered, large false positive rates persist even asymptotically (in T). When the hypotheses are well separated, as in the case $\gamma = 20$, then this distortion reduces, but since the power of the DF test increases, there is little gain (if any) in using either the I-SPSM or the C-SPMS, especially in the presence of relatively small panels ($N \leq 20$).

The observed behaviour is not related to the cumulation of small size distortions described in Hanck (2008): rather, it is the result of the interaction of the power of the panel UR test with the probability that a small p value be associated with a test under the alternative, as anticipated in Section 2. When the local alternative is close to the null, there is a non-trivial probability that some of the small p values do not correspond to tests under the alternative. When γ increases, the null and the alternative hypotheses become more clearly separated and a better classification can be attained. However, in this case correctly specified conventional DF tests have approximately the same tpr with greater precision than that obtained using either the I-SPSM or the D-SPSM.

The overall classifiers' behaviour synthesized in Figure 3, while consistent with the implications outlined in Section 2, contradicts Chortareas & Kapetanios's statement that

> "An ideal situation for SPSM is one where most series considered are stationary and very persistent." (Chortareas & Kapetanios, 2009, p. 393)

The other procedures remain very conservative. In fact, the H-SPSM and the related closed testing procedure (Hommel, 1988) never reach a better result than a half of the true positive rate that can be attained using conventional DF tests, with only a very limited increase of precision.

Dependent test statistics

As is well known, the panel UR tests proposed by Choi (2001) and Im *et al.* (2003) are valid only in the presence of independent tests statistics. Therefore, in the presence of positively dependent test statistics we expect high *fpr*'s as far as the I-SPSM procedure is concerned. Furthermore, we substitute the C-SPSM with the D-SPSM, based on Demetrescu *et al.* (2006), in order to take into account dependence in the p value combination test. The H-SPSM remains valid also in the presence of positively dependent test statistics (see Hanck, 2013). To save space, in this Section we report only the simulation results relative to the case $\rho = 0.8$ and $N_1/N \in \{0.2, 0.8\}$: the results with $\rho = 0.4$ and those with $N_1/N = 0.5$ are qualitatively similar and are available in the supplementary material in the appendix.

When the fraction of stationary time series is small (see Figure 4), apart from the expected large values of the *fpr* for the I-SPSM, the results are qualitatively similar to those obtained under independence (Figure 2). However, the true positive rate and precision of the panel-based classification procedures are even lower than those obtained in the presence of independent tests.

Compared to the case of independent tests, when the fraction of stationary series is large (see Figure 5) the excess of positives disappears, except again for the I-SPSM. This is due to the fact that dependence makes the p values under the null and under the alternative more clearly separated, so that a smaller number of null hypotheses are wrongly

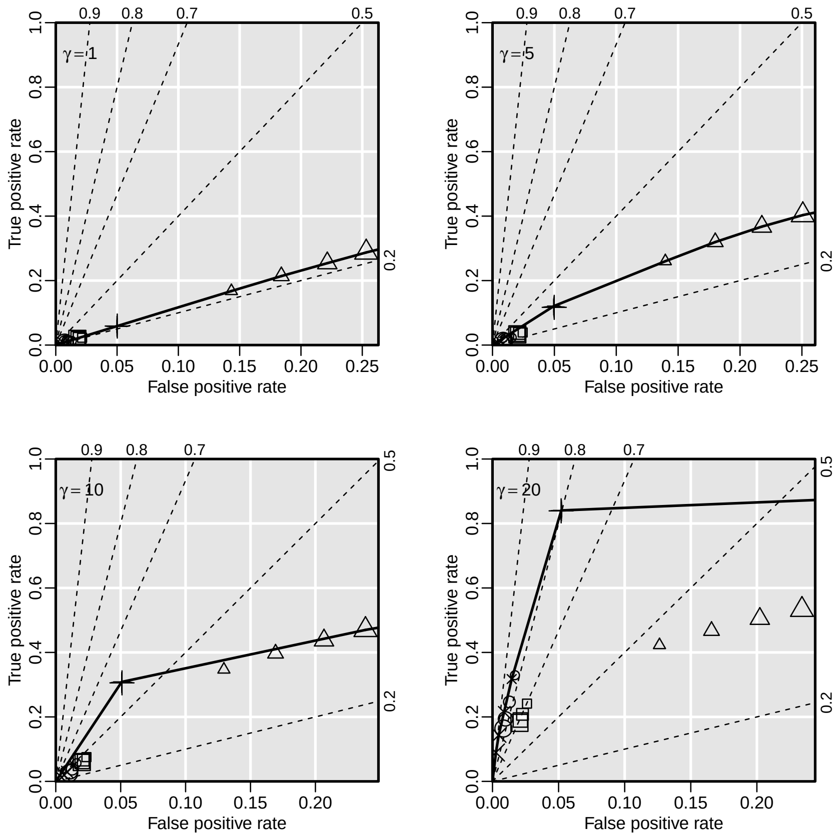

Figure 4. Dependent test statistics ($\rho = 0.8$). 5,000 replications with the fraction of stationary series $N_1/N = 0.2$. The local-to-unit root alternative is $\varrho = \exp(-\gamma/T)$ with $\gamma \in \{1, 5, 10, 20\}$. Larger symbols correspond to larger panels with $N \in \{10, 20, 40, 80\}$. The dashed lines are precision isometrics. The lowest precision isometric coincides with the random guessing line. The broken solid line is the ROC convex hull. $\square$ = D-SPSM; $\triangle$ = I-SPSM; $\bigcirc$ = H-SPSM; $+$ = DF; $\times$ = Hommel.

classified as alternative. However, also in this case the panel-based classification methods have in general worse classification performances than those obtained by using standard DF tests directly. Only in one case (when $\gamma = 20$) the D-SPSM gets the same true positive rate and better precision than the DF. Despite this, the precision increment is so small that it is legitimate to ask if it justifies the extra conceptual and computational burden involved in using the D-SPSM.

5. Conclusions

Panel unit root tests are powerful tools, but rejection of the joint unit root null hypothesis does not convey information about which series in a panel are stationary. In order to overcome this difficulty, Chortareas & Kapetanios (2009) suggested using a recursive panel testing procedure, which they labelled Sequential Panel Selection Method.

This paper investigates the relative merits of this procedure as compared to the standard time series approach based on individual Dickey-Fuller tests. In particular, the classification performance of four adaptations of the Sequential Panel Selection Method are considered, the first being the original one proposed by Chortareas & Kapetanios (2009), which is based on the panel unit root test proposed in Im *et al.* (2003); the second and third are based on the p value combination tests suggested in Choi (2001) and Demetrescu *et al.* (2006), respectively; the fourth version is based on the intersection test advo-

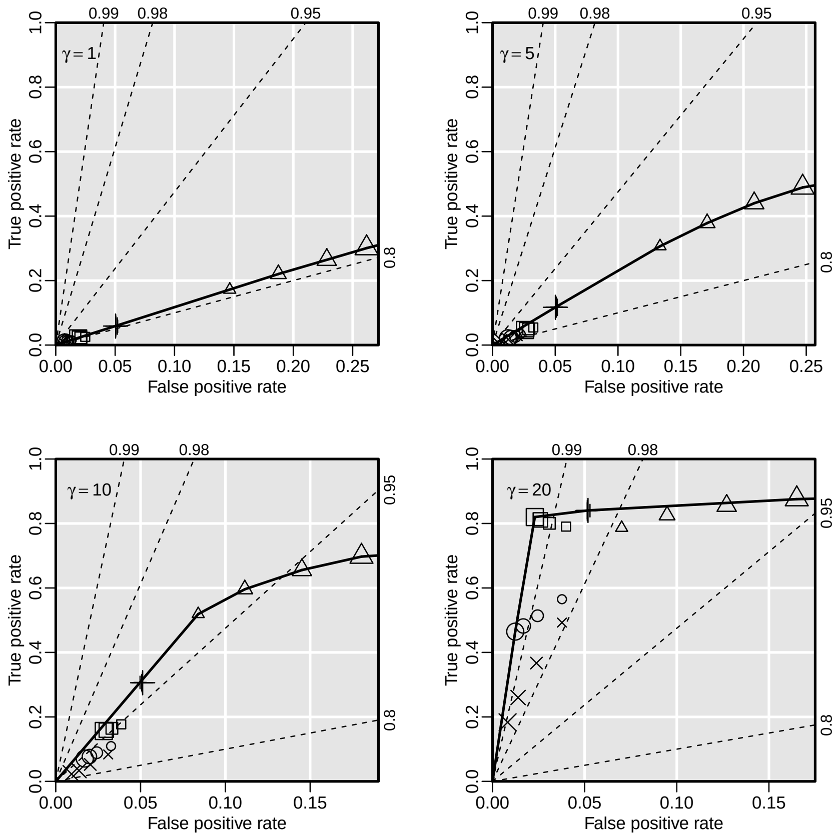

Figure 5. Dependent test statistics ($\rho = 0.8$). 5,000 replications with the fraction of stationary series $N_1/N = 0.8$. The local-to-unit root alternative is $\varrho = \exp(-\gamma/T)$ with $\gamma \in \{1, 5, 10, 20\}$. Larger symbols correspond to larger panels with $N \in \{10, 20, 40, 80\}$. The dashed lines are precision isometrics. The lowest precision isometric coincides with the random guessing line. The broken solid line is the ROC convex hull. $\square$ = D-SPSM; $\triangle$ = I-SPSM; $\bigcirc$ = H-SPSM; $+$ = DF; $\times$ = Hommel.

cated in Hanck (2013). The method suggested in Hommel (1988), which is directly related to Hanck's (2013), is also considered.

In order to focus on the relative merits of the classification procedures and not on the finite sample properties of the underlying individual unit root tests, the analysis is based on the direct simulation of the asymptotic test t statistics and p values, either under the null or under selected local-to-unit root alternatives. This can be seen equivalent to analyzing the classification ability of the different procedures under the best conditions, where no specification errors in the Dickey-Fuller equations or finite sample biases take place.

Our results show that the investigated procedures offer no special advantages to appropriately used time series-based UR tests. Furthermore, the results confirm that the panel-based classification procedures are sensitive to the composition of the panel in terms of the fraction of time series under the null and under the alternative and to the existence and strength of dependence across individual tests. The only classification procedure which in all Monte Carlo experiments lies on the convex hull is the one based on the individual UR tests. Therefore, we should conclude that the time series-based tests are the most efficient classification tools among those considered here. Given the nature of our Monte Carlo analysis, the general conclusion should remain unchanged if more powerful individual unit root tests and/or different panel UR tests are used.

References

Chan NH (1988). "The Parameter Inference for Nearly Nonstationary Time Series." *Journal of the American Statistical Association*, **83**(403), 857–862.

Choi I (2001). "Unit Root Tests for Panel Data." *Journal of International Money and Finance*, **20**(2), 249–272.

Chortareas G, Kapetanios G (2009). "Getting PPP Right: Identifying Mean-reverting Real Exchange Rates in Panels." *Journal of Banking and Finance*, **33**(2), 390–404.

Costantini M, Lupi C (2013). "A Simple Panel-CADF Test for Unit Roots." *Oxford Bulletin of Economics and Statistics*, **75**(2), 276–296.

Demetrescu M, Hassler U, Tarcolea AI (2006). "Combining Significance of Correlated Statistics with Application to Panel Data." *Oxford Bulletin of Economics and Statistics*, **68**(5), 647–663.

Fawcett T (2006). "An Introduction to ROC Analysis." *Pattern Recognition Letters*, **27**(8), 861–874.

Flach PA (2003). "The Geometry of ROC Space: Understanding Machine Learning Metrics through ROC Isometrics." In T Fawcett, N Mishra (eds.), *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*, pp. 194–201. AAAI Press, Washington DC.

Hanck C (2008). "The Error-in-Rejection Probability of Meta-Analytic Panel Tests." *Economics Letters*, **101**(1), 27–30.

Hanck C (2013). "An Intersection Test for Panel Unit Roots." *Econometric Reviews*, **32**(2), 183–203.

Hatanaka M (1996). *Time-Series-Based Econometrics: Unit Roots and Cointegration*. Advanced Texts in Econometrics. Oxford University Press, Oxford.

Hommel G (1988). "A Stagewise Rejective Multiple Test Procedure Based on a Modified Bonferroni Test." *Biometrika*, **75**(2), 383–386.

Im KS, Pesaran MH, Shin Y (2003). "Testing for Unit Roots in Heterogeneous Panels." *Journal of Econometrics*, **115**(1), 53–74.

Maddala GS, Wu S (1999). "A Comparative Study of Unit Root Tests with Panel Data and a New Simple Test." *Oxford Bulletin of Economics and Statistics*, **61**(Supplement 1), 631–652.

O'Connell PGJ (1998). "The Overvaluation of Purchasing Power Parity." *Journal of International Economics*, **44**(1), 1–19.

Phillips PCB (1987). "Towards a Unified Asymptotic Theory for Autoregression." *Biometrika*, **74**(3), 535–547.

Provost F, Fawcett T (2001). "Robust Classification for Imprecise Environments." *Machine Learning*, **42**(3), 203–231.

R Development Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://R-project.org`.

Sarkar SK, Chang CK (1997). "The Simes Method for Multiple Hypothesis Testing With Positively Dependent Test Statistics." *Journal of the American Statistical Association*, **92**(440), 1601–1608.

Simes RJ (1986). "An Improved Bonferroni Procedure for Multiple Tests of Significance." *Biometrika*, **73**(3), 751–754.

Appendix A: Supplementary material

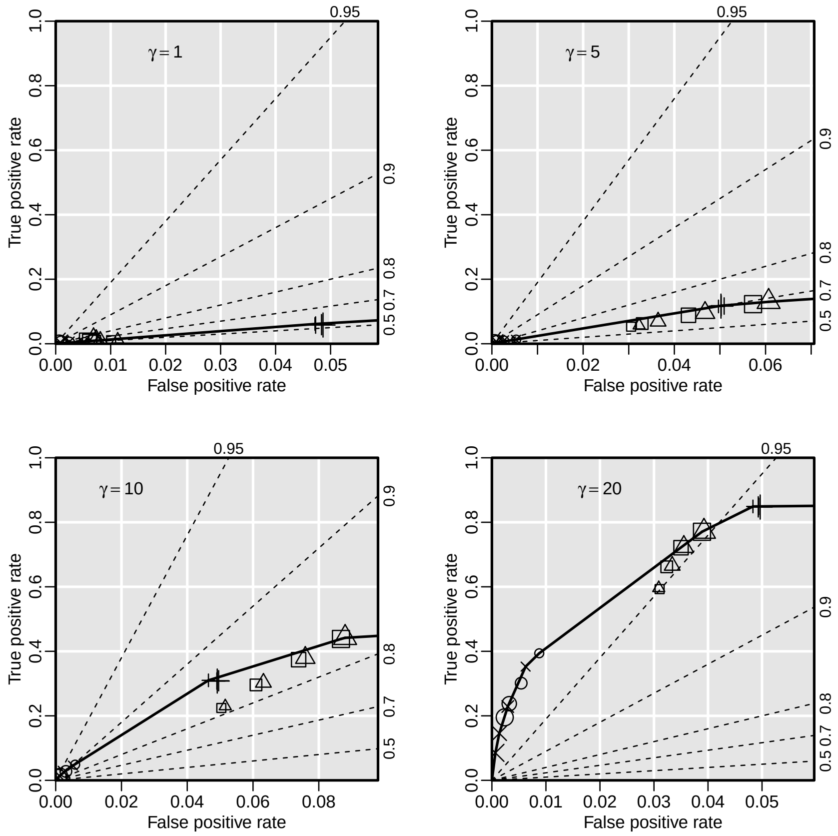

Figure 6. Independent test statistics. 5,000 replications with the fraction of stationary series $N_1/N = 0.5$. The local-to-unit root alternative is $\varrho = \exp(-\gamma/T)$ with $\gamma \in \{1, 5, 10, 20\}$. Larger symbols correspond to larger panels with $N \in \{10, 20, 40, 80\}$. The dashed lines are precision isometrics. The lowest precision isometric coincides with the random guessing line. The broken solid line is the ROC convex hull. $\square$ = C-SPSM; $\triangle$ = I-SPSM; $\bigcirc$ = H-SPSM; $+$ = DF; $\times$ = Hommel.

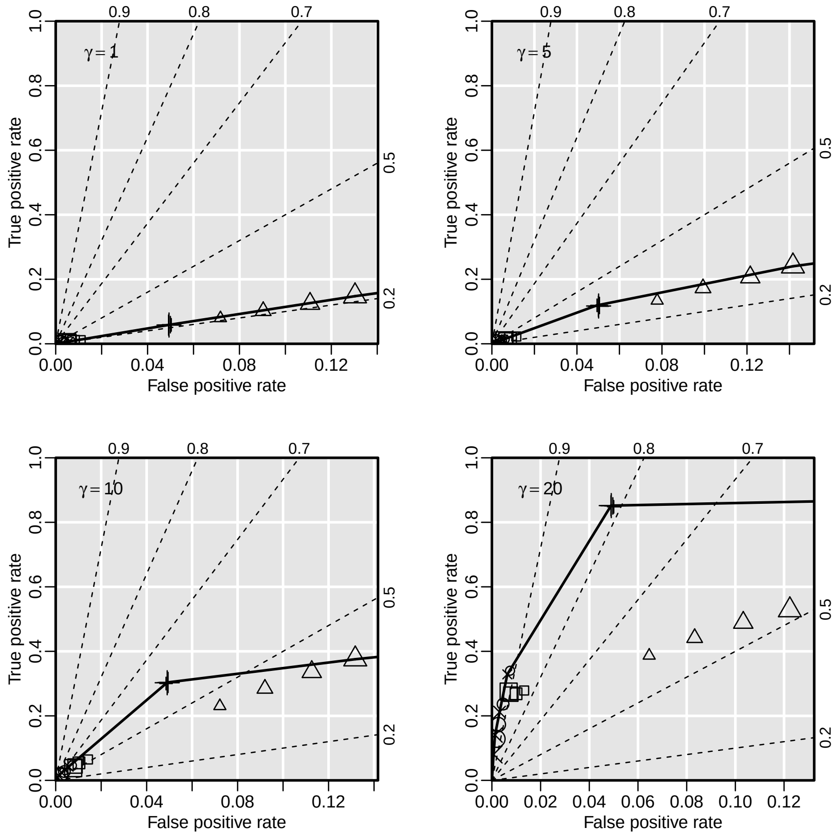

Figure 7. Dependent test statistics ($\rho = 0.4$). 5,000 replications with the fraction of stationary series $N_1/N = 0.2$. The local-to-unit root alternative is $\varrho = \exp(-\gamma/T)$ with $\gamma \in \{1, 5, 10, 20\}$. Larger symbols correspond to larger panels with $N \in \{10, 20, 40, 80\}$. The dashed lines are precision isometrics. The lowest precision isometric coincides with the random guessing line. The broken solid line is the ROC convex hull. $\square$ = D-SPSM; $\triangle$ = I-SPSM; $\bigcirc$ = H-SPSM; $+$ = DF; $\times$ = Hommel.

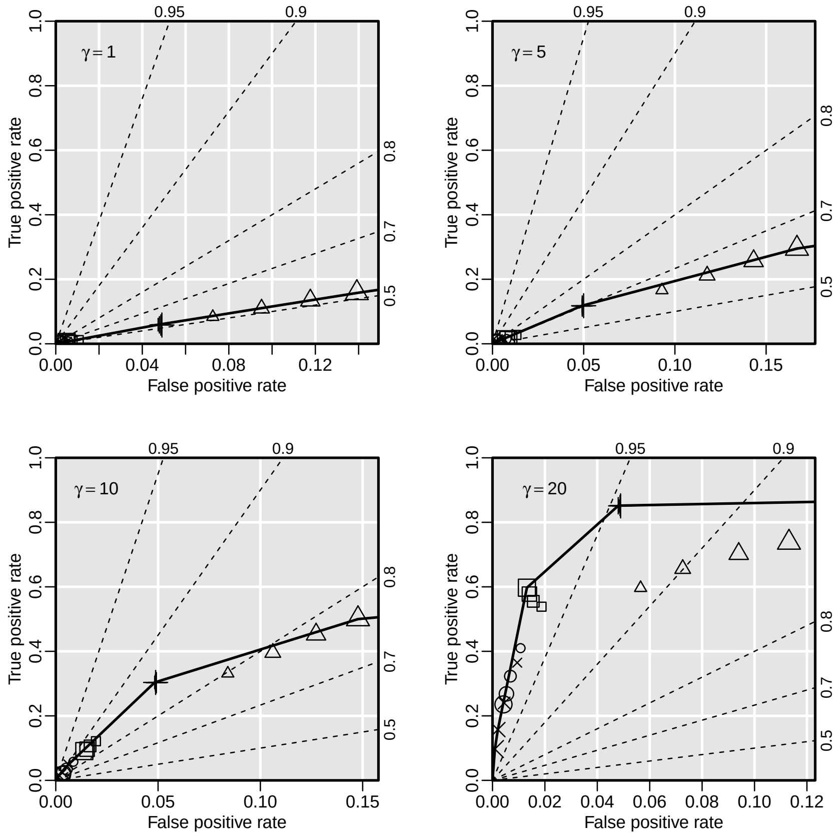

Figure 8. Dependent test statistics ($\rho = 0.4$). 5,000 replications with the fraction of stationary series $N_1/N = 0.5$. The local-to-unit root alternative is $\varrho = \exp(-\gamma/T)$ with $\gamma \in \{1, 5, 10, 20\}$. Larger symbols correspond to larger panels with $N \in \{10, 20, 40, 80\}$. The dashed lines are precision isometrics. The lowest precision isometric coincides with the random guessing line. The broken solid line is the ROC convex hull. $\square$ = D-SPSM; $\triangle$ = I-SPSM; $\bigcirc$ = H-SPSM; $+$ = DF; $\times$ = Hommel.

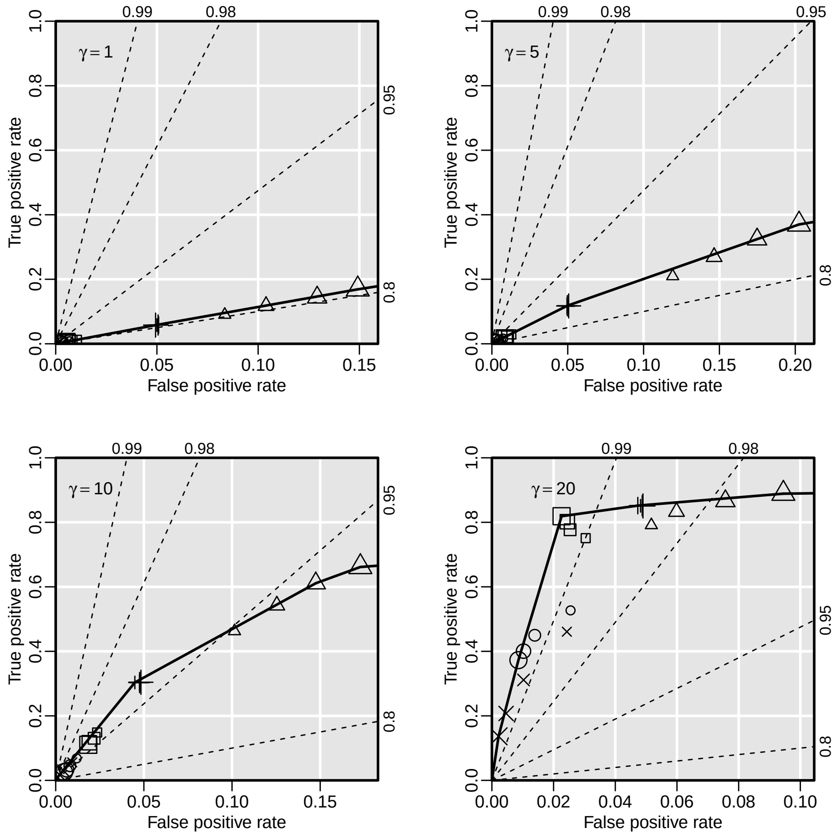

Figure 9. Dependent test statistics ($\rho = 0.4$). 5,000 replications with the fraction of stationary series $N_1/N = 0.8$. The local-to-unit root alternative is $\varrho = \exp(-\gamma/T)$ with $\gamma \in \{1, 5, 10, 20\}$. Larger symbols correspond to larger panels with $N \in \{10, 20, 40, 80\}$. The dashed lines are precision isometrics. The lowest precision isometric coincides with the random guessing line. The broken solid line is the ROC convex hull. $\square$ = D-SPSM; $\triangle$ = I-SPSM; $\bigcirc$ = H-SPSM; $+$ = DF; $\times$ = Hommel.

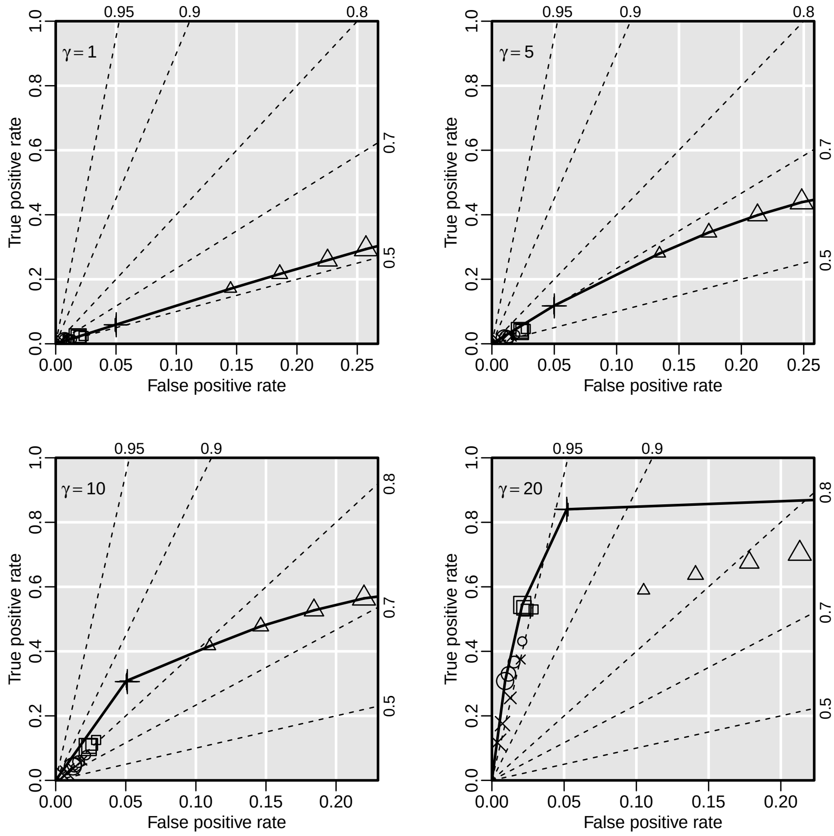

Figure 10. Dependent test statistics ($\rho = 0.8$). 5,000 replications with the fraction of stationary series $N_1/N = 0.5$. The local-to-unit root alternative is $\varrho = \exp(-\gamma/T)$ with $\gamma \in \{1, 5, 10, 20\}$. Larger symbols correspond to larger panels with $N \in \{10, 20, 40, 80\}$. The dashed lines are precision isometrics. The lowest precision isometric coincides with the random guessing line. The broken solid line is the ROC convex hull. $\square$ = D-SPSM; $\triangle$ = I-SPSM; $\bigcirc$ = H-SPSM; $+$ = DF; $\times$ = Hommel.